A Deep Denoising Autoencoder Approach to Improving the Intelligibility of Vocoded Speech in Cochlear Implant Simulation

Ying-Hui Lai, Fei Chen, *Member, IEEE,* Syu-Siang Wang, Xugang Lu, Yu Tsao*, *Member, IEEE*,
and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*—Objective: In a cochlear implant (CI) speech processor, noise reduction (NR) is a critical component for enabling CI users to attain improved speech perception under noisy conditions. Identifying an effective NR approach has long been a key topic in CI research. Method: Recently, a deep denoising autoencoder (DDAE) based NR approach was proposed and shown to be effective in restoring clean speech from noisy observations. It was also shown that DDAE could provide better performance than several existing NR methods in standardized objective evaluations. Following this success with normal speech, the present work further investigated the performance of DDAE-based NR to improve the intelligibility of envelope-based vocoded speech, which simulates speech signal processing in existing CI devices. Results: We compared the performance of speech intelligibility between DDAE-based NR and conventional single-microphone NR approaches using the noise vocoder simulation. The results of both objective evaluations and listening test showed that, under the conditions of nonstationary noise distortion, DDAE-based NR yielded higher intelligibility scores than conventional NR approaches. Conclusion and Significance: This study confirmed that DDAE-based NR could potentially be integrated into a CI processor to provide more benefits to CI users under noisy conditions.**

***Index Terms*—cochlear implant, noise reduction, deep denoising autoencoder, vocoder simulation.**

I. Introduction

COCHLEAR implants (CIs) are surgically implanted electronic devices that provide a sense of sound in patients with profound-to-severe hearing loss. The considerable progress of CI technologies in the past three decades has enabled many CI users to enjoy a high level of speech understanding in quiet. For most CI users, however, understanding speech in noisy environments remains challenging [1-3]. Various noise reduction (NR) methods have been developed and implemented in CI processors to improve speech perception under diverse noisy conditions [4-6]. They can be broadly divided into multiple- and single-microphone NR approaches. Multi-microphone approaches are beneficial when the target and noise are spatially separated. The direction of arrival for a sound source is exploited to spatially filter the signal and remove the noise. Spriet et al. [7] evaluated a two-microphone adaptive beamformer, BEAM, in the Nucleus Freedom CI system for speech understanding with background noise. The results confirmed that the approach might significantly increase the speech perception capabilities of CI users under noisy conditions. Hersbach et al. [6] tested a combination of NR approaches designed to improve CI performance in noise; based on the signal-to-noise ratio (SNR) estimation, the performance was evaluated in combination with several directional microphone approaches available in the Cochlear CP810 sound processor. The results indicated that multi-microphone directionality was effective in improving speech understanding under spatially separated noisy conditions. Subsequently, Hersbach et al. [8] proposed a beamformer postfilter, in which the noise was spatially separated from the target speech so as to improve the performance for CI users under noisy conditions. More recently, Buechner et al. [9] investigated the performance of monaural and binaural beamforming technologies with an additional NR approach for CI users. Their study showed that both adaptive and binaural beamformers were significantly superior to omni-directional microphones for CI users.

Although multi-microphone methods can achieve satisfactory performances in intelligibility [6-12], a secondary microphone and headphone combination is required; this increases the hardware cost (both microphone and battery are necessary). In addition, the performance of multi-microphone methods may degrade in reverberant environments, and their applicability is restricted to acoustic situations in which the target speech and noise are spatially separated [13]. Compared to multi-microphone methods, single-microphone NR methods are aesthetically more appealing and economically more feasible [3]. Conventional single-microphone NR methods have been adopted for CI processors. Successful examples include log minimum mean squared error (logMMSE) [14], Karhunen-Loéve transform (KLT) [15, 16], Wiener filter based on a priori SNR estimation (Wiener) [17], ClearVoice NR [5], and SNR-based [4] approaches. Most of these methods have focused on designing a filter by exploring the statistical distributions of speech and noise signals. Improved performance can be guaranteed when noise and speech are separable in an explored space. Recently, Chen et al. [3] evaluated the aforementioned NR methods for Mandarin-speaking CI users and found that

This work was supported by the National Natural Science Foundation of China (Grant No. 61571213).

Ying-Hui Lai is with the Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan (e-mail:yhlai@ee.yzu.edu.tw).

Yu Tsao and Syu-Siang Wang are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan (e-mail: yu.tsao@citi.sinica.edu.tw and sypdbhee@gmail.com).

Fei Chen is with Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China (e-mail: fchen@sustc.edu.cn)

Xugang Lu is with National Institute of Information and Communications Technology, Japan (e-mail: xugang.lu@nict.go.jp)

Chin-Hui Lee is with School of Electrical and Computer Engineering, Georgia Institute of Technology, GA, USA (e-mail: chl@ece.gatech.edu)

although most single-microphone NR approaches effectively improved CI speech recognition in noise, they performed differently in various environmental noises. Furthermore, NR methods should be tailored to individual type of noise for the CI processor to gain the optimal cost/benefit tradeoff. Moreover, most of these single-microphone NR methods perform according to either the additive nature of the statistical properties of the speech and noise signals or the additive nature of the background noise. However, they typically fail to track nonstationary noise for real-world scenarios in which acoustic conditions are unknown [18]. In other words, though single-microphone NR approaches can provide significant benefits to CI users, their performance can still be improved.

Recently, deep learning-based NR approaches have been proposed and confirmed to be effective in various NR tasks [18-24]. Its nonlinear processing can characterize high order statistical information accurately and thus be used for NR. It is believed that a deep network (i.e., multiple hidden layers) is preferable to a shallow network (i.e., with single or few hidden layers). Among these approaches, Lu et al. [20, 21] proposed a deep denoising autoencoder (DDAE)-based NR approach that involves converting noisy speech into clean speech through a series of nonlinear transformations. In implementation, a DDAE model is trained to encode statistical information pertaining only to clean speech to transform noisy speech into clean speech. Through this processing, the DDAE model explicitly learns the statistical differences between clean and noisy speech. Previous studies have confirmed that the performance of the DDAE-based NR approach is superior to that of conventional single-microphone NR approaches (e.g., MMSE plus an improved minimum controlled recursive averaging noise-tracking algorithm [25]) for normal hearing (NH) according to several standardized objective evaluations [20]. However, the performance of the DDAE-based NR approach for CI speech processing remains unknown.

The aim of the present study is to evaluate the performance of the DDAE-based NR approach on the basis of speech intelligibility for vocoded speech (which simulates the speech signal processing normally used in a CI device [26]) under various noisy conditions. We compare the performance of speech recognition between DDAE-based NR and conventional single-microphone NR methods. In this study, we intend to investigate the noise suppression capability of NR methods in challenging conditions. Two nonstationary noises at lower SNR levels were used to form the test data for evaluation. We first use an objective evaluation (i.e., short-time objective intelligibility measure (STOI) [27]) to confirm the effectiveness of DDAE-based NR on normal speech. The STOI score has been shown high correlation with speech intelligibility for normal speech [27]. However, the correlation of STOI and intelligibility of vocoded speech has not been confirmed, and a previous study indicated that the normalized covariance measure (NCM) [28] can be used to predict the intelligibility of vocoded Mandarin speech [29]. Therefore, we adopted the NCM to evaluate the speech intelligibility performance of vocoded speech processed by DDAE-based NR and conventional NR approaches. Subsequently, listening experiment involving NH subjects and vocoded speech was conducted to further evaluate the performance in human subjects. In addition, processed envelopes are used to qualitatively analyze the advantage of DDAE-based NR over conventional NR approaches.

The rest of this paper is organized as follows: Section II introduces conventional NR approaches and DDAE-based NR. Section III presents the vocoder-based speech synthesis process. Section IV demonstrates the experimental setup and results. Section V gives concluding remarks.

II. Noise Reduction

The goal of NR is to reduce noise components from noisy speech to generate enhanced speech with improved SNR, intelligibility, and perceptual quality. Assuming that $\boldsymbol{y}$, $\boldsymbol{x}$, and $\boldsymbol{n}$, denote noisy, clean, and noise signals in the time domain, respectively, we have

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{n}. \tag{1}$$

In the spectral domain, the noisy speech spectrum, $\boldsymbol{Y}_l$, can be expressed as

$$\boldsymbol{Y}_l = \boldsymbol{X}_l + \boldsymbol{N}_l\,, \tag{2}$$

where $\boldsymbol{X}_l$ and $\boldsymbol{N}_l$ are the speech and noise spectra of the l-th frequency bin, respectively, corresponding to frequency ω_l, where $\omega_l = 2\pi l/L, l = 0, 1, \ldots, L-1$. The aim of NR approaches is to restore $\boldsymbol{x}$ (or $\boldsymbol{X}_l$) from $\boldsymbol{y}$ (or $\boldsymbol{Y}_l$).

Various NR approaches have been proposed. A notable class of NR approaches is spectral restoration. This class of approaches aims to estimate a gain function, $\boldsymbol{G}_l$, based on the statistics of speech and noise. The enhanced speech, $\widehat{\boldsymbol{X}}_l$, is obtained by filtering $\boldsymbol{Y}_l$ through $\boldsymbol{G}_l$. The phase of the noisy speech is borrowed and used to prepare the phase of the clean speech. An inverse FFT (IFFT) is applied to convert $\widehat{\boldsymbol{X}}_l$ and the phase to obtain enhanced speech $\widehat{\boldsymbol{x}}$. Notable spectral restoration approaches include spectral subtraction [30] and Wiener filter [17] with their various extensions [31-33]. Some spectral restoration approaches are derived from probabilistic models of speech and noise signals. Well-known approaches include maximum a posteriori spectral amplitude estimator [34], maximum likelihood spectral amplitude estimator [35], generalized MAPA [36], minimum mean-square-error (MMSE) spectral estimator [37], and logMMSE [14].

Another successful class of NR approaches is the subspace method. This class of approaches adopts a linear estimator, $\mathbf{H}$ (a $K \times K$ matrix), to obtain enhanced speech, $\widehat{\boldsymbol{x}}$, by:

$$\widehat{\boldsymbol{x}} = \mathbf{H} \cdot \boldsymbol{x} + \mathbf{H} \cdot \boldsymbol{n}. \tag{3}$$

The error signal $\boldsymbol{\varepsilon}$ is estimated by:

$$\boldsymbol{\varepsilon} = \widehat{\boldsymbol{x}} - \boldsymbol{x} = (\mathbf{H} - \mathbf{I}) \cdot \boldsymbol{x} + \mathbf{H} \cdot \boldsymbol{n} = \boldsymbol{\varepsilon}_{\boldsymbol{X}} + \boldsymbol{\varepsilon}_{\boldsymbol{n}}, \tag{4}$$

where $\boldsymbol{\varepsilon}$ can be divided into $\boldsymbol{\varepsilon}_{\boldsymbol{X}}$ and $\boldsymbol{\varepsilon}_{\boldsymbol{n}}$, which represent the speech distortion and residual noise, respectively. The transformation, $\mathbf{H}$, is estimated by minimizing the speech distortion with the constraint of a predetermined level of residual noise.

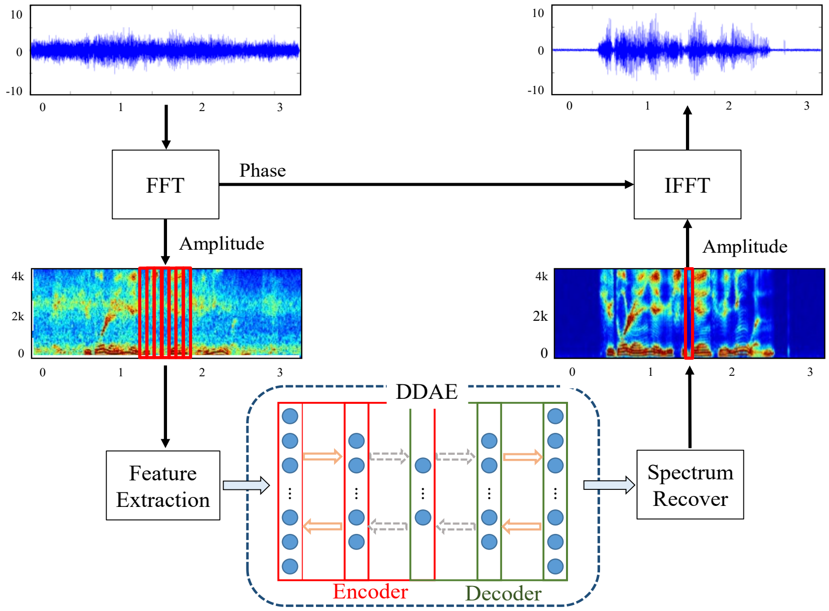

Fig. 1. Structure of a DDAE-based NR system.

Originally, the subspace NR approach is derived to handle the white Gaussian noise (WGN) [38]. Extensions of the subspace NR approaches have been proposed to handle colored noise [15, 16].

More recently, deep learning-based NR approaches have attracted great attention [18-21]. With the multiple layers of nonlinear processing, the deep learning NR approaches can accurately characterize the mapping function from noisy to clean speech signals, and have been confirmed to achieve outstanding performance in various NR tasks. Among the deep learning-based NR approaches, this study focuses on the DDAE-based NR approach. The DDAE model has been widely used to design deep neural architectures for robust feature extraction and classification [39]. Lu et al. proposed the use of DAE and its deep variant, DDAE, to perform NR [20]. Fig. 1 shows the structure of the DDAE-based NR approach.

The DDAE-based NR procedure can be divided into training and testing phases. In training, a set of noisy-clean speech pairs is prepared. The noisy-clean speech signals are first converted into the frequency domain by an FFT. The logarithm amplitudes of noisy and clean speech spectra are then placed in the input and output sides of the DDAE model, respectively. More specifically, the input is a vector containing logarithm amplitudes of the noisy spectrum:

$$\boldsymbol{Y}_m^E = \left[\log\left(\left|Y_{1,m-\tau}\right|\right) \ldots \log\left(\left|Y_{L,m-\tau}\right|\right)\right] \ldots \log\left(\left|Y_{1,m}\right|\right), \ldots \log(|Y_{l,m}|) \ldots \log\left(\left|Y_{L,m}\right|\right) \ldots \log\left(\left|Y_{1,m+\tau}\right|\right) \ldots \log\left(\left|Y_{L,m+\tau}\right|\right)]' \quad ,$$

where τ is the window length to characterize the context information; the output is a vector of logarithm amplitudes, $\boldsymbol{X}_m^E = [\log(|X_{1,m}|) \ldots \log(|X_{l,m}|) \ldots \log\,(|X_{L,m}|)]$ of the clean speech spectrum, where $|Y_{l,m}|$ and $|X_{l,m}|$ are the amplitudes of the noisy and clean spectra, respectively at the l-th frequency bin and the m-th frame. For a DDAE model with J hidden layers, we have

$$\begin{gathered} h^1(\boldsymbol{Y}_m^E) = \sigma(\boldsymbol{W}^1\boldsymbol{Y}_m^E + \boldsymbol{b}^1), \\ \vdots \\ h^J(\boldsymbol{Y}_m^E) = \sigma(\boldsymbol{W}^{J-1}h^{J-1}(\boldsymbol{Y}_m^E) + \boldsymbol{b}^{J-1}), \\ \widehat{\boldsymbol{X}}_m^E = \boldsymbol{W}^J h^J(\boldsymbol{Y}_m^E) + \boldsymbol{b}^J, \end{gathered} \tag{5}$$

where $\{\boldsymbol{W}^1 \ldots \boldsymbol{W}^J\}$ are the matrices of the connection weights, $\{\boldsymbol{b}^1 \ldots \boldsymbol{b}^J\}$ are the bias vectors, and $\widehat{\boldsymbol{X}}_m^E$ is the vector containing logarithm amplitudes of restored speech corresponding to the noisy counterpart $\boldsymbol{Y}_m^E$. The nonlinear function $\sigma(.)$ of a hidden neuron is a logistic function defined as

$$\sigma(t) = 1/(1 + exp(-t)). \tag{6}$$

The parameters are determined by optimizing the following objective function:

$$\begin{gathered} \theta^* = \arg\min_\theta(F(\theta) + \eta^1\|\boldsymbol{W}^1\|_\mathrm{F}^2 + \cdots + \\ \eta^L\|\boldsymbol{W}^L\|_\mathrm{F}^2), \\ F(\theta) = \frac{1}{M}\sum_{m=1}^{M}\left\|\boldsymbol{X}_m^E - \widehat{\boldsymbol{X}}_m^E\right\|_2^2, \end{gathered} \tag{7}$$

where $\theta = \{\boldsymbol{W}^1 \ldots \boldsymbol{W}^J; \boldsymbol{b}^1 \ldots \boldsymbol{b}^J\}$ is the parameter set of the DDAE model, and M is the total number of training samples. In Eq. (7), $\{\eta^1 \ldots \eta^L\}$ controls the tradeoff between the reconstruction accuracy and regularization of the weighting coefficients (we set $\eta^1 = \cdots = \eta^L$=0.0002 in this study), and $\|.\|_\mathrm{F}^2$ denotes the Frobenius norm. Eq. (7) can be optimized using any unconstrained optimization algorithm. In this study, we used a Hessian-free algorithm [40] to compute the parameters, θ. In the testing phase, the logarithm amplitudes of noisy speech signals

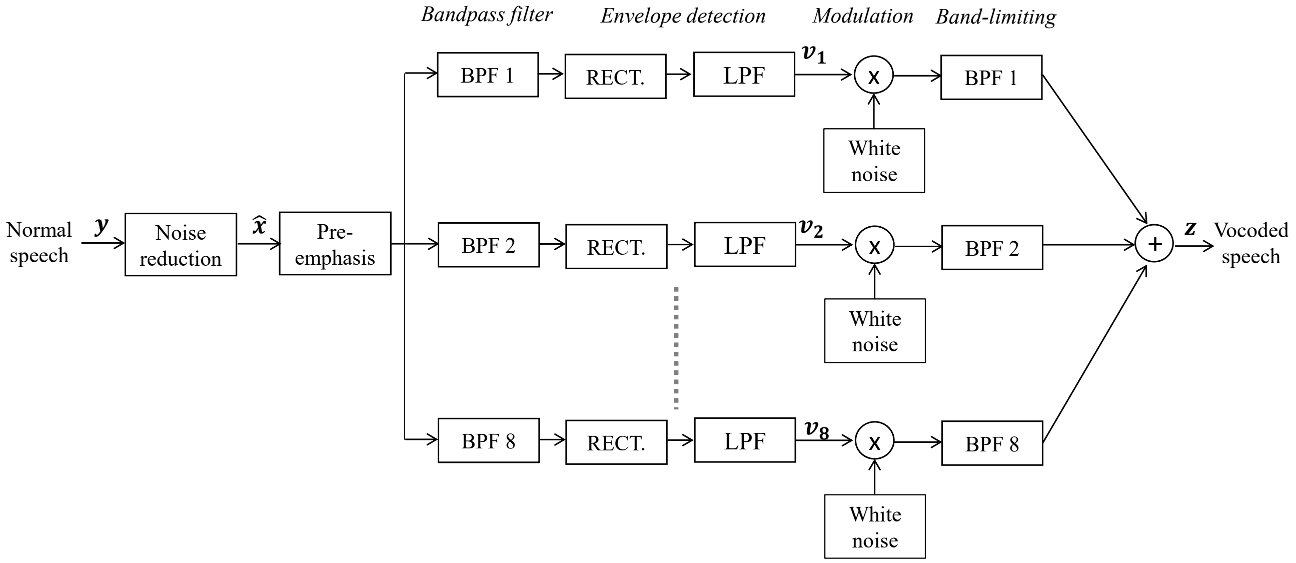

Fig. 2. Block diagram of an eight-channel noise-vocoder.

are inputted to the trained DDAE model to obtain the logarithm amplitudes of enhanced speech signals for the output. Similar to the spectral restoration approaches, the phases of the noisy speech are borrowed to prepare the phases for the enhanced speech. With the DDAE-enhanced amplitudes and the phase information, the enhanced speech can be synthesized. More detailed information on DDAE-based NR can be found in [20, 21].

III. Vocoded Speech With Envelope

Vocoder (Voice Operated reCOrDER) is a voice processing system that can analyze and resynthesize human voice signals. The first vocoder was developed by Homer Dudley from Bell Laboratories [1] and introduced to the public at the 1939–1940 New York World's Fair [2]. After being invented, the vocoder has been used in a wide variety of applications, including audio data compression, voice encryption and transmission, and voice modification. In past decades, vocoder has had a profound impact on the development of CI research; that is, vocoder-based speech simulations have been widely adopted to predict the general pattern of speech recognition performance for CI users [26, 41-44]. For such simulations, speech signals are processed by a vocoder to mimic the sounds heard by CI patients. The sounds are then presented to NH subjects for listening tests. There are two advantages of using vocoder simulations to predict CI users' speech perception: (1) it overcomes the difficulty of conducting experiments on real CI recipients, especially for some particular regions in which most recipients are young children [44]; and (2) it avoids the impact of patient-specific confounding factors (e.g., neural surviving pattern) that exist in clinical populations [29, 41, 45]. Numerous studies have confirmed that vocoder simulations could predict the pattern of listening performance for CI users, including the effects of background noise [48], type of speech masker [49], and number of electrodes [3, 50, 51]. It is noted that vocoder simulations are not expected to predict the absolute performance level of each CI user but rather the trend in performance when a particular parameter is varied. The present study also conducted speech recognition experiments using vocoder simulations and NH subjects.

There are two main types of vocoder implementations for CI experiments—namely, tone vocoder and noise vocoder [46]. The difference between these two types of vocoders lies in the modulation step. For a tone vocoder, sinewaves at the center frequencies of band-pass filters are used as carriers to modulate the baseband signals, and the output generated is a sum of the amplitude-modulated sinewaves from different channels [42]. For a noise vocoder (see Fig. 2), on the other hand, white noise is used as carrier, and the output is a sum of amplitude-modulated noise (i.e., with band-limiting processing) from different channels [26, 47]. Dorman et al. [42] compared English speech intelligibility using a tone vocoder and a noise vocoder with a varying number of channels; their results indicated that the two types of vocoders showed only small differences, not reaching statistical significance in most test conditions with vowels, consonants, and sentences. The noise vocoder has been used in many recent studies, and this study continues to adopt noise vocoders to simulate CI speech processing [29].

Fig. 2 shows the block diagram of our simulation system, which consists of an eight-channel noise-vocoder and an NR stage. In Fig. 2, the normal speech signals, $\boldsymbol{y}$, are first processed by the NR stage to generate enhanced signals, $\hat{\boldsymbol{x}}$. The signals are processed by the noise vocoder, which comprises the pre-emphasis stage, band-pass filter (BPF) stage, envelope detection stage, modulation stage, and band-limiting stage. The pre-emphasis stage involves a 3 dB/octave roll-off filter with a cutoff frequency of 2000 Hz. The BPFs then filter the emphasized signals into eight frequency bands between 80 and 6000 Hz (with cutoff frequencies at 80, 221, 426, 724, 1158, 1790, 2710, 4050, and 6000 Hz). The temporal envelope v_k in k-th band is extracted using a full-wave rectifier (i.e., the RECT module in Fig. 2) followed by a low-pass filter (LPF) with a 400-Hz cutoff frequency. The envelopes for all bands are then modulated using a set of white noise, and further filtered using the same set of BPFs. Finally, the modulated sinewaves of the eight bands are summed together, and the level of the combined signal is

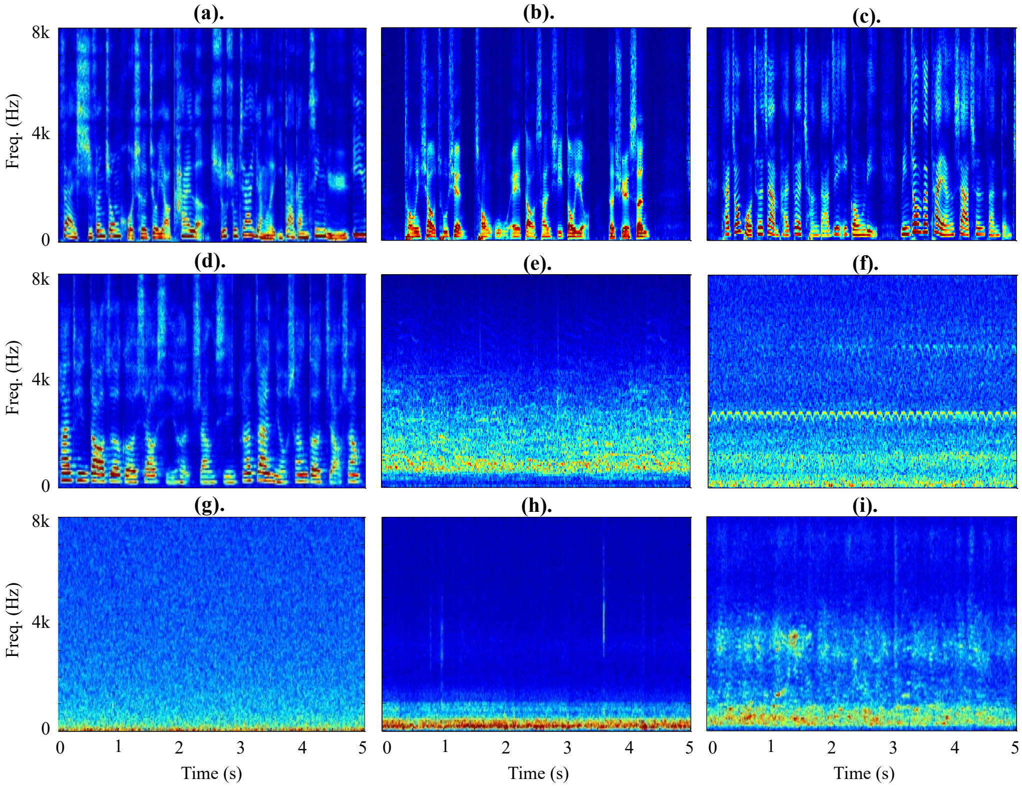

Fig. 3. Spectrograms of noise signals of (a) man1, (b) man2, (c) woman1, (d) woman2, (e) crowd cheering, (f) buccaneer1, (g) pink, (h) car, and (i) cafeteria babble. These nine noises were used to synthesize noisy speech for the training data.

adjusted to produce a root-mean-square value equal to that of the original input wideband signal.

IV. Experimental Results and Discussion

In this study, we evaluated the capability of the DDAE-based NR to improve the intelligibility of speech in mismatched conditions (i.e., different noises and speech utterances were used at the training and testing phases). The performances of conventional NR methods, including logMMSE, KLT, and Wiener techniques, which have been shown to benefit the intelligibility for CI recipients [3], were also tested for comparison. For the DDAE-based NR, 280 clean utterances (approximately 2.5 sec. each, and different from those used in the testing phase) of fluent Mandarin Chinese speech were used for training. Nine noise types were adopted, including speech from four Mandarin speakers (i.e., man1, man2, woman1, and woman2); crowd cheering, buccaneer1, and pink noises [18]; and car and cafeteria babble noises [48]. Fig. 3 shows the spectrograms of these nine noises. These nine noises were artificially added to the clean training utterances to generate −10, −5, 0, 5, and 10 dB signals. A total of 12,600 (280 × 9 × 5) utterances were formed as the training set. As mentioned earlier, the log-scale power spectrum coefficients were used as the acoustic features for the training set. The features were extracted from a 16 ms windowed signal with an 8 ms frame shift. The DDAE model has five layers, with 500 neurons in each layer.

We conducted three sets of experiments—two objective evaluations and one listening test. For the objective tests, we adopted STOI to evaluate normal wideband speech processed by the four NR approaches (signal $\hat{\boldsymbol{x}}$ in Fig. 2). Next, we used NCM to estimate the intelligibility scores of the four NR-processed vocoded speech (signal $\boldsymbol{z}$ in Fig. 2). For the listening test, we measured the sentence recognition results of the four NR approaches by presenting vocoded speech (signal $\boldsymbol{z}$ in Fig. 2) to NH individuals. All three experiments were conducted using the eight-channel noise vocoder as shown in Fig. 2.

A. *Objective Evaluation*

The objective evaluations were performed using speech excerpts from the Mandarin Chinese version of Hearing in Noise Test (MHINT) [49]. All utterances were pronounced by a male native speaker, with a fundamental pitch frequency ranging from 75 to 180 Hz and recorded at a sampling rate of 16 kHz.

Fig. 4 shows the noise spectrograms of a cocktail party and two equal-level interfering female talkers (2T), which were used to prepare two challenging listening conditions for the CI recipients [50, 51]. A total of 50 clean test utterances excerpted from the MHINT corpus were corrupted by cocktail noise [48] and 2T [44] maskers at −12, −9, −6, −3, 0, 3, 6, 9, and 12dB SNR levels to form the testing set. 18 test conditions (9 SNR levels × 2 noise environments) with a total of 900 testing noisy utterances were set to evaluate the performance of logMMSE-, KLT-, Wiener-, and DDAE-based NR techniques. Notably, we considered the two following points when designing the training and testing sets: (1) the DDAE-based NR did not acquire prior knowledge about the noise signals online because the noise types used in the training and testing data were different, which enabled a fairer comparison of the DDAE-based and conventional NR methods; and (2) we intended to focus more on challenging conditions, and, thus, evaluated the performances using two nonstationary noises at low SNR conditions.

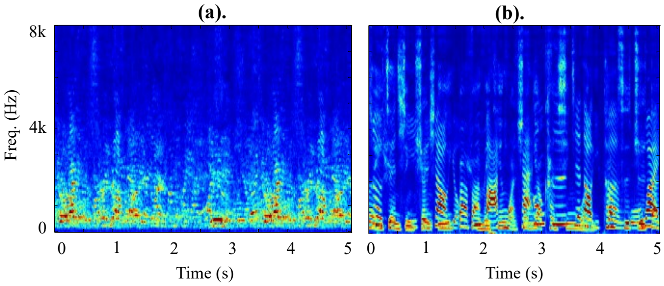

Fig. 4. Spectrograms of noise signals of (a) cocktail noise and (b) 2T masker. These two noises were used to synthesize noisy speech for the testing data.

1) *Objective Evaluation on Normal Speech*

In the previous studies, DDAE-based NR has been confirmed to provide a better perceptual estimation of speech quality (PESQ) and lower restoration error scores than conventional NR approaches [20, 21]. However, whether it can also effectively enhance speech intelligibility has not been verified. In this section, we intend to evaluate the intelligibility scores of the DDAE-based NR-processed speech. The standardized STOI measure is used for evaluation.

The STOI measure is derived based on a correlation coefficient between the temporal envelops of the clean and degraded (or processed) speech in short-time regions. Previous studies have showed that STOI is highly relevant to human speech intelligibility [27]. The STOI score ranges from 0 to 1; a higher score corresponds to a better speech intelligibility result. There are four steps to compute an STOI score: (1) DFT-based 1/3 octave band decomposition, which is used to convert a speech waveform to a representation that resembles the transform properties of the human auditory system; (2) short-time segmentation, which divides each entire signal into several short-time overlapping segments; (3) normalization and clipping, which compensate for global level differences that should not have a strong effect on the speech intelligibility; and (4) score computation, which estimates correlation coefficients of 1/3 octave bands and finally an average intelligibility score by averaging all bands and frames. During score computation, the silence regions are excluded. More details about the four steps of the STOI measure can be found in [27].

Figs. 5(a) and (b) show the average STOI scores at nine different SNR conditions for the cocktail and 2T maskers, respectively. For the results of cocktail masker in Fig. 5(a), the average STOI scores for {noisy, logMMSE, KLT, Wiener, DDAE} are {0.48, 0.36, 0.29, 0.36, 0.53} at -12 dB SNR, {0.54, 0.42, 0.37, 0.43, 0.61} at -9dB SNR, {0.61 0.49, 0.45, 0.51, 0.67} at -6dB SNR, {0.68, 0.56, 0.53, 0.58, 0.73} at -3 dB SNR, {0.75, 0.62, 0.61, 0.64, 0.77} at 0 dB SNR, {0.80, 0.68, 0.67, 0.70, 0.80} at 3dB SNR, {0.85, 0.72, 0.72, 0.74, 0.82} at 6 dB SNR, {0.89, 0.76, 0.76, 0.79, 0.84} at 9 dB SNR, and {0.92, 0.79, 0.78, 0.82, 0.85} at 12 dB SNR. For the 2T masker results in Fig. 5(b), the average STOI scores for {noisy, logMMSE, KLT, Wiener, DDAE} are {0.41, 0.31, 0.25, 0.31, 0.48} at -12 dB SNR, {0.47, 0.35, 0.29, 0.35, 0.54} at -9dB SNR, {0.55, 0.44, 0.40, 0.44, 0.61} at -6dB SNR, {0.64, 0.53, 0.50, 0.54, 0.68} at -3 dB SNR, {0.71, 0.60, 0.58, 0.61, 0.75} at 0 dB SNR, {0.77, 0.66, 0.64, 0.68, 0.80} at 3dB SNR, {0.83, 0.71, 0.70, 0.73, 0.85} at 6 dB SNR, {0.87, 0.75, 0.74, 0.77, 0.89} at 9 dB SNR, and {0.91, 0.78, 0.77, 0.80, 0.92} at 12 dB SNR.

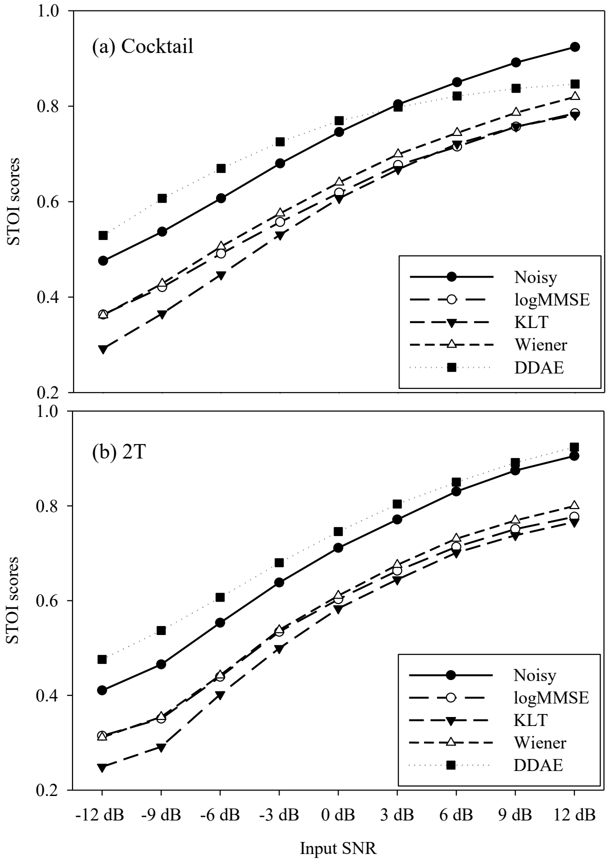

Fig. 5. Average STOI scores at nine SNR conditions for (a) cocktail noise and (b) 2T masker.

From the results of Fig. 5, DDAE consistently achieves better STOI scores than logMMSE, KLT and Wiener for both cocktail and 2T maskers over the nine different SNR conditions. The results of STOI indicated that DDAE provides large improvements over the other conventional NR algorithms, because such techniques actually decrease the performance substantially when speech is in the nonstationary noise conditions. Moreover, DDAE can provide clearer gains over logMMSE, KLT and Wiener for 2T than cocktail masker condition, as shown in Fig. 5(b). There is a wider gap between DDAE and the three conventional KLT, logMMSE and Wiener NR techniques than that in Fig. 5(a), which shows a narrower gap, suggesting that DDAE provides more benefits for competing masker types. In addition, the results indicated that the conventional NR techniques obtained lower STOI score than that of noisy speech in all test conditions; on the other hand, DDAE obtained higher STOI scores than noisy speech in most test conditions. Notably, the DDAE provides a higher intelligibility improvement from the baselines on low SNRs than those on high SNRs. This finding again suggests that the DDAE NR systems are robust in more challenging noisy environments.

2) *Objective Evaluation on Vocoded Speech*

In this section, we present the results of the intelligibility evaluation for vocoded speech. The NCM measure was used in this set of experiments. The NCM is a speech transmission index (TI)-related measurement [52], which is estimated based on the covariance of the envelopes between the clean and processed signals. Previous studies have shown that the NCM measure can correlate well with the intelligibility of vocoded speech, mainly due to the similarities of the NCM calculation and CI processing strategies; both use information extracted from the envelope in a number of frequency bands while discarding fine-structure information [53]. There are four steps to compute an NCM score: (1) a set of BPFs was used to decompose the clean and processed speech into *N* bands; (2) the envelope of each band was extracted using the Hilbert transform and then down-sampled to $2f_{\text{cut}}$ Hz, thereby limiting the envelope modulation rate to f_{cut} Hz; (3) the normalized correlation coefficient and SNR values of each band are computed, and the TI value in each band are computed by linearly mapping the SNR values to the range between 0 and 1; (4) the NCM score is obtained by averaging the TI values across all frequency bands with a set of weights. In this study, the ANSI articulation index weights were used as the coefficients [54], and f_{cut} and N were set to 200 and 20, respectively. The NCM score ranges from 0 to 1; a higher score corresponds to better speech intelligibility. More detail about the NCM measure can be found in [28, 55].

Figs. 6(a) and (b) show the average NCM scores at nine different SNR conditions for the cocktail and 2T maskers, respectively. For the cocktail masker results in Fig. 6(a), the average NCM scores for the {noisy, logMMSE, KLT, Wiener, DDAE} are {0.01, 0.02, 0.02, 0.01, 0.11} at -12 dB SNR, {0.02, 0.03, 0.04, 0.03, 0.15} at -9dB SNR, {0.03, 0.03, 0.04, 0.03, 0.18} at -6dB SNR, {0.06, 0.05, 0.07, 0.06, 0.20} at -3 dB SNR, {0.12, 0.09, 0.13, 0.11, 0.21} at 0 dB SNR, {0.18, 0.14, 0.19, 0.15, 0.23} at 3dB SNR, {0.22, 0.18, 0.24, 0.20, 0.24} at 6 dB SNR, {0.25, 0.21, 0.26, 0.23, 0.24} at 9 dB SNR, and {0.26, 0.23, 0.26, 0.24, 0.25} at 12 dB SNR. For the 2T masker results in Fig. 4(b), the average NCM scores for {noisy, logMMSE, KLT, Wiener, DDAE} are {0.02, 0.02, 0.02, 0.02, 0.10} at -12 dB SNR, {0.01, 0.02, 0.02, 0.02, 0.13} at -9dB SNR, {0.05, 0.04, 0.05, 0.04, 0.17} at -6dB SNR, {0.08, 0.07, 0.09, 0.07, 0.20} at -3 dB SNR, {0.13, 0.11, 0.14, 0.12, 0.21} at 0 dB SNR, {0.19, 0.17, 0.20, 0.18, 0.23} at 3dB SNR, {0.22, 0.20, 0.24, 0.21, 0.24} at 6 dB SNR, {0.24, 0.22, 0.25, 0.23, 0.25} at 9 dB SNR, and {0.25, 0.23, 0.26, 0.24, 0.24} at 12 dB SNR.

From the NCM results in Fig. 6, DDAE outperforms the other three NR approaches on vocoded speech consistently over the seven SNR levels (i.e., -12 to 6 dB) in both cocktail and 2T conditions. Although the NCM scores are not as good as the STOI scores in Fig. 5(a) and (b), the gap between DDAE and the other NR approaches in Fig. 6 for the NCM scores are wider than that of the STOI scores in Fig. 5, indicating that the improvement gains of DDAE are more obvious in vocoded speech conditions. These two sets of results are consistent with the STOI results of normal speech in Fig. 5, again confirming the effectiveness of DDAE's NR capability in challenging (low SNR and nonstationary noise) listening conditions.

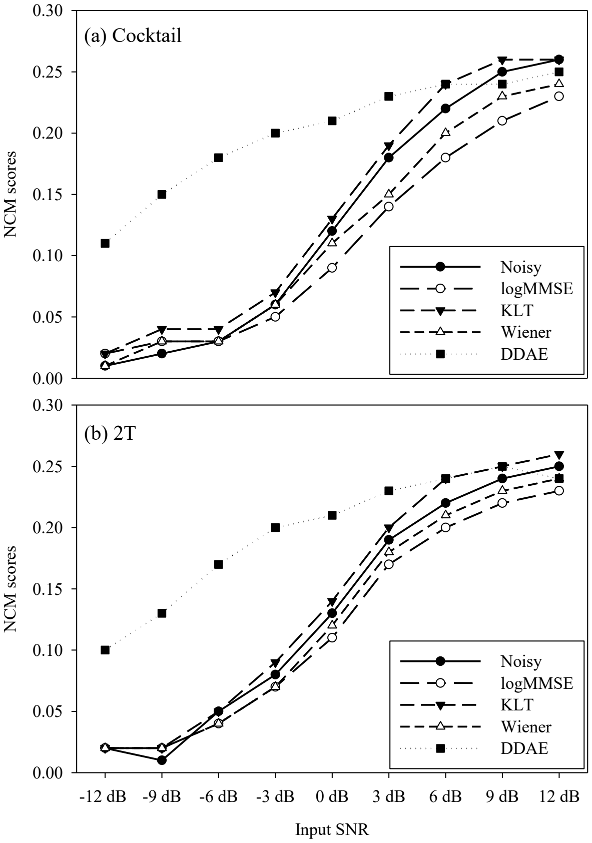

Fig. 6. Average NCM scores at nine SNR conditions for (a) the cocktail masker and (b) the 2T masker.

B. *Listening Test*

This section reports the listening test results of noisy, logMMSE, KLT, Wiener and DDAE with real subjects. Ten NH native Mandarin Chinese subjects (seven males and three females) aged 18–22 participated in the listening tests. The same MHINT sentences used in the objective evaluations were adopted in the listening tests. The cocktail and 2T maskers were again used to corrupt the test sentences. Because real subjects were involved in this set of experiments, the number of test sets is confined to avoid biased results caused by listening fatigue [56] and ceiling effects of speech recognition [57]. Thus, we decided to prepare only two SNR levels (i.e., 0 and 3 dB) based on the results of objective evaluation for each masker.

The experiments were conducted in a soundproof booth. The stimuli were played to the subjects through a set of Sennheiser HD headphones at a comfortable listening level. Each subject participated in a total of 20 test conditions: 2 SNR levels × 2 maskers × 5 NR techniques—i.e., noisy, logMMSE, KLT, Wiener and DDAE. Each condition contained ten sentences, and the order of the 20 conditions was randomized individually for each listener. None of the ten sentences was repeated across the test conditions. The subjects were instructed to verbally repeat what they heard and were allowed to repeat the stimuli twice. The word correct rate (WCR) is used as the evaluation metric, which is calculated by dividing the number of correctly identified words by the total number of words under each test

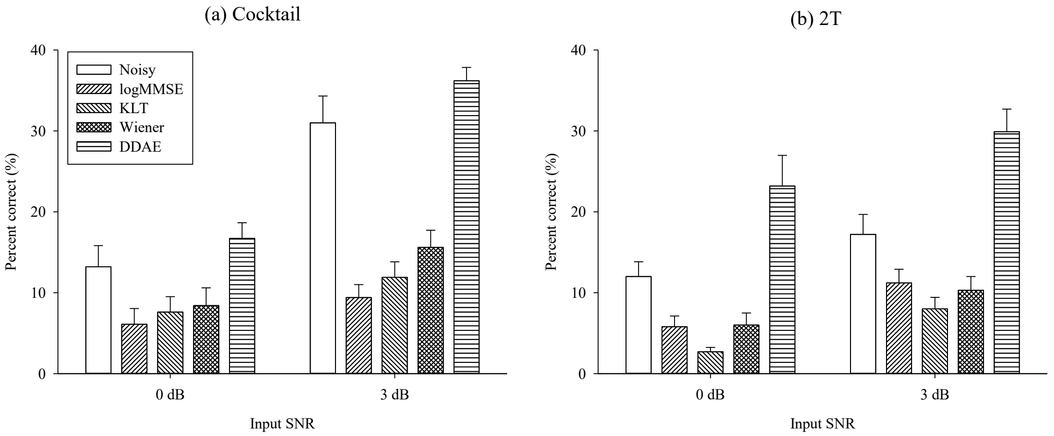

Fig. 7. Average speech recognition scores in 0 and 3 dB SNR conditions for (a) the cocktail masker and (b) the 2T masker. Each error bar indicates one standard error of the mean (SEM).

condition. During the test, each subject was given a 5-min break after every 30 min of testing.

Fig. 7(a) and (b) show the average WCRs for the cocktail and 2T maskers, respectively, at different SNR conditions. For the cocktail masker results in Fig. 7 (a), the average±SEM WCRs for {noisy, logMMSE, KLT, Wiener, DDAE} are {13.2±2.6, 6.1±1.9, 7.6±1.9, 8.7±2.2, 16.7±1.9} at 0 dB SNR and {31.0±3.3, 9.4±1.6, 11.9±1.9, 15.6±2.1, 36.2±1.6} at 3 dB SNR. For the 2T masker results in Fig. 7 (b), the average±SEM WCRs for {noisy, logMMSE, KLT, Wiener, DDAE} are {12±1.8, 5.8±1.3, 2.7±0.5, 6±1.5, 23.2±3.7} at 0 dB SNR, and {17.2±2.5, 11.2±1.7, 8±1.4, 10.3±1.7, 29.9±2.8} at 3 dB SNR.

Because of the floor effect, the scores were first converted to rational arcsine units (RAU) using the rationalized arcsine transform [58]. Table 1 presents the results of the one-way analysis of variance (ANOVA) and Duncan post-hoc comparisons on four test conditions (i.e., cocktail and 2T noise types with 0 and 3 dB SNRs) for the noisy, logMMSE, KLT, Wiener, and DDAE (denoted as N, L, K, W, and D in the table, respectively) techniques. The mean scores in the table are equal to the scores of "percent correct" in Fig. 7. The ANOVA results for the cocktail masker confirm that the speech recognition scores are significantly different over the five groups, with (p = 0.05) and (p < 0.001) on 0 and 3 dB SNRs, respectively. The Duncan post-hoc comparisons further verify the significant differences for the following group pairs: (noisy, logMMSE), (DDAE, logMMSE), (DDAE, KLT), and (DDAE, Wiener) at 0 dB SNR; (noisy, logMMSE), (noisy, KLT), (noisy, Wiener), (KLT, logMMSE), (KLT, Wiener), (DDAE, logMMSE), (DDAE, KLT), and (DDAE, Wiener) at 3 dB SNR. Furthermore, the ANOVA results for the 2T masker confirm that the speech recognition scores significantly differ across the five groups, with p < 0.001 at 0 and 3 dB SNRs, respectively. The Duncan post-hoc comparisons further verify the significant difference for the following group pairs at both 0 and 3 dB SNRs: (noisy, logMMSE), (noisy, KLT), (noisy, Wiener), (DDAE, noisy), (DDAE, logMMSE), (DDAE, KLT), and (DDAE, Wiener).

Table 1. One-way ANOVA statistical analysis test and Duncan post-hoc testing in four test conditions and five NR testing sets.

Test conditions	**Processed**	**Mean**	***p***	**Post-hoc comparison**
Cocktail (0dB)	N	13.2	0.05	(N, L), (D, L)
	L	6.1		(D, K), (D, W)
	K	7.6		
	W	8.4		
	D	16.7		
Cocktail (3dB)	N	31	<0.001	(N, L), (N, K)
	L	9.4		(N, W), (K, L)
	K	11.9		(K, W), (D, L)
	W	15.6		(D, K), (D, W)
	D	36.2		
2T (0dB)	N	12	<0.001	(N, L), (N, K)
	L	5.8		(N, W), (D, N)
	K	2.7		(D, L), (D, K)
	W	6		(D, W)
	D	23.2		
2T (3dB)	N	17.2	<0.001	(N, K), (N, W)
	L	11.2		(N,W), (D,N)
	K	8		(D,L), (D,K)
	W	10.3		(D,W)
	D	29.9		

Note: N, L, K, W and D are the testing sets of noisy, logMMSE, KLT, Wiener and DDAE, respectively.

Results of the listening tests in Fig. 7 and Table 1 were consistent with the STOI and NCM results in Figs. 5 and 6, providing the following four observations: (1) DDAE outperforms logMMSE, KLT and Wiener for cocktail and 2T at both 0dB and 3 dB SNRs; (2) DDAE provides more gains than the other three NR approaches in these four test conditions—namely, 0 and 3 dB for both cocktail and 2T; (3) DDAE provides remarkable gains than the other three NR approaches in the nonstationary masker; and (4) interestingly, the performance of mean score for noisy speech is often comparable to or better than the three conventional NR-processed speech signals but is always worse than the DDAE-processed signals. The ANOVA and Duncan post-hoc comparisons further confirm that DDAE significantly higher than noisy speech in 2T test conditions. The above observations confirm the DDAE's superior properties in improving the intelligibility in challenging conditions. Because the vocoder simulation could indicate the relative performance of different NR methods, the results suggest that DDAE can be potentially integrated into a CI processor to improve speech perception under noisy conditions.

Previous studies confirmed that conventional NR methods can provide satisfactory performance in stationary or relatively slow nonstationary noise [16, 59]. We focused our attention herein on NR methods for more challenging tasks involving nonstationary noises (i.e., cocktail party and 2T maskers), which CI recipients may encounter in real-life conditions. Fig. 7 shows that conventional NR methods yielded a poorer speech intelligibility scores than that of the original noisy speech. The worse speech intelligibility performance may be obtained from an inaccurate noise estimation (i.e., noise tracking) process performed to compute noise statistics [36, 60]. Annoying residual noise or speech distortion may occur if the noise statistics are not accurately estimated, thereby resulting in poor speech intelligibility. Numerous noise estimation algorithms, such as voice activity detection [61], minima-controlled recursive algorithm [62], Martin [63], and minima-controlled recursive algorithm version 2 (MCRA2), have been proposed over the past two decades [60]. These noise estimation algorithms provide a satisfactory NR performance against stationary noises. However, most of them could not perform very well in challenging conditions (e.g., 2T) [60]. The above reasons may explain the unsatisfactory performance achieved by the conventional NR methods under nonstationary conditions, especially when the background noise is competing speech signals, shown in Fig. 7.

The fundamental mechanism of DDAE acts similarly to the companding strategies [64-66] that based on signal processing to achieve a spectral enhancement from noisy speech. More specifically, the companding strategy was directly inspired by the mechanism of auditory processing for acoustic signals, i.e., the two-tone suppression effect (or the lateral inhibition effect but without any lateral coupling between channels). This effect helps to improve the spectral contrast which is important in enhancing spectral peaks in noisy environments. The DDAE algorithm however works with a different starting point, i.e., a machine learning approach to examining the difference of the statistics between speech and noise, and emphasize the speech property (or structure) in noisy environment. As it is known that the speech signal shows strong regular structures, e.g., spectral formants, temporal modulation, etc., in this sense, the DDAE algorithm also tries to enhance the spectral contrast and temporal modulation by discriminative learning between speech and noise. In short, DDAE and the bio-inspired companding strategy try to achieve the same target but with different starting points to help CI recipients to improve the speech intelligibility performance. DDAE can be regarded as a data-driven supervised learning method that has the advantage of automatically obtaining the nonlinear transform function between noisy and clean speech.

V. Conclusion

This study investigated the effectiveness of the DDAE-based NR in improving the intelligibility of vocoded speech simulating CI speech processing. Based on the results of the objective evaluations and listening tests, the DDAE-based NR provided superior performance in terms of speech intelligibility compared to three conventional NR approaches (i.e., logMMSE, KLT, and Wiener). The contribution of the paper is two-fold. First, we confirmed that the DDAE-based NR could provide higher intelligibility scores for envelope-based vocoded speech in challenging conditions (nonstationary noise at low SNR levels) when compared to conventional NR methods. Second, the results suggested that deep learning-based NR methods could potentially be implemented in CI speech processors to provide benefits to CI recipients.

The present work also conducted objective NCM evaluations and listening tests by using a noise-vocoded speech simulation. Such simulation is widely used because a newly designed approach for CI processing can be more easily evaluated by NH listeners. Furthermore, subject-dependent factors associated with hearing loss and neural surviving patterns can be bypassed. Many studies have successfully predicted outcomes by using such tone-vocoder simulations (e.g., [42, 51, 55, 67-71]). Notably, some inconsistences could occur when transferring the designed approach to real CI devices. Therefore, although we have confirmed the capability of the DDAE-based NR to provide higher intelligibility than conventional approaches, a further evaluation need be conducted to confirm the effectiveness of the DDAE-NR based on speech intelligibility tests with real CI recipients in appropriate clinical situations.

References

[1] P. Loizou, "Speech processing in vocoder-centric cochlear implants." pp. 109-143.

[2] B. L. Fetterman, and E. H. Domico, “Speech recognition in background noise of cochlear implant patients,” *Otolaryngology--Head and Neck Surgery,* vol. 126, no. 3, pp. 257-263, 2002.

[3] F. Chen, Y. Hu, and M. Yuan, “Evaluation of Noise Reduction Methods for Sentence Recognition by Mandarin-Speaking Cochlear Implant Listeners,” *Ear and hearing,* vol. 36, no. 1, pp. 61-71, 2015.

[4] P. W. Dawson, S. J. Mauger, and A. A. Hersbach, “Clinical evaluation of signal-to-noise ratio–based noise reduction in Nucleus® cochlear implant recipients,” *Ear and hearing,* vol. 32, no. 3, pp. 382-390, 2011.

[5] A. Buechner, M. Brendel, H. Saalfeld, L. Litvak, C. Frohne-Buechner, and T. Lenarz, “Results of a pilot study with a signal enhancement algorithm for HiRes 120 cochlear implant users,” *Otology & Neurotology,* vol. 31, no. 9, pp. 1386-1390, 2010.

[6] A. A. Hersbach, K. Arora, S. J. Mauger, and P. W. Dawson, “Combining directional microphone and single-channel noise reduction algorithms: a clinical evaluation in difficult listening conditions with cochlear implant users,” *Ear and hearing,* vol. 33, no. 4, pp. 13-23, 2012.

[7] A. Spriet, L. Van Deun, K. Eftaxiadis, J. Laneau, M. Moonen, B. van Dijk, A. Van Wieringen, and J. Wouters, “Speech understanding in background noise with the two-microphone adaptive beamformer BEAM™ in the Nucleus Freedom™ Cochlear Implant System,” *Ear and Hearing,* vol. 28, no. 1, pp. 62-72, 2007.

[8] A. A. Hersbach, D. B. Grayden, J. B. Fallon, and H. J. McDermott, “A beamformer post-filter for cochlear implant noise reductiona,” *The Journal of the Acoustical Society of America,* vol. 133, no. 4, pp. 2412-2420, 2013.

[9] A. Buechner, K. H. Dyballa, P. Hehrmann, S. Fredelake, and T. Lenarz, “Advanced Beamformers for Cochlear Implant Users: Acute Measurement of Speech Perception in Challenging Listening Conditions,” *PloS one,* vol. 9, no. 4, pp. e95542, 2014.

[10] R. Van Hoesel, and G. M. Clark, “Evaluation of a portable two-microphone adaptive beamforming speech processor with cochlear implant patients,” *The Journal of the Acoustical Society of America,* vol. 97, no. 4, pp. 2498-2503, 1995.

[11] V. Hamacher, W. Doering, G. Mauer, H. Fleischmann, and J. Hennecke, “Evaluation of noise reduction systems for cochlear implant users in different acoustic environment,” *American Journal of Otology,* vol. 18, no. 6, pp. 46-49, 1997.

[12] K. Kokkinakis, and P. C. Loizou, “Using blind source separation techniques to improve speech recognition in bilateral cochlear implant patients,” *The Journal of the Acoustical Society of America,* vol. 123, no. 4, pp. 2379-2390, 2008.

[13] J. Wouters, and J. V. Berghe, “Speech recognition in noise for cochlear implantees with a two-microphone monaural adaptive noise reduction system,” *Ear and hearing,* vol. 22, no. 5, pp. 420-430, 2001.

[14] Y. Ephraim, and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 33, no. 2, pp. 443-445, 1985.

[15] A. Rezayee, and S. Gazor, “An adaptive KLT approach for speech enhancement,” *Speech and Audio Processing, IEEE Transactions on,* vol. 9, no. 2, pp. 87-95, 2001.

[16] Y. Hu, and P. C. Loizou, “A generalized subspace approach for enhancing speech corrupted by colored noise,” *Speech and Audio Processing, IEEE Transactions on,* vol. 11, no. 4, pp. 334-341, 2003.

[17] P. Scalart, “Speech enhancement based on a priori signal to noise estimation,” *in Proc. ICASSP,* vol. 2, pp. 629-633, 1996.

[18] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, “A Regression Approach to Speech Enhancement Based on Deep Neural Networks,” *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 23, no. 1, pp. 7-19, 2015.

[19] Y. Xu, J. Du, L. Dai, and C. Lee, “An experimental study on speech enhancement based on deep neural networks,” *Signal Processing Letters, IEEE,* vol. 21, no. 1, pp. 65-68, 2014.

[20] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” *in Proc. Interspeech*, pp. 436-440, 2013.

[21] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Ensemble modeling of denoising autoencoder for speech spectrum restoration,” *in Proc. Interspeech*, pp. 885-889, 2014.

[22] Y. Wang, A. Narayanan, and D. Wang, “On training targets for supervised speech separation,” *Audio, Speech, and Language Processing, IEEE/ACM Transactions on,* vol. 22, no. 12, pp. 1849-1858, 2014.

[23] A. Narayanan, and D. Wang, “Ideal ratio mask estimation using deep neural networks for robust speech recognition,” *in Proc. ICASSP*, pp. 7092-7096, 2013.

[24] S. W. Fu, Y. Tsao, and X. Lu, “SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement,” *Interspeech 2016*, pp. 3768-3772, 2016.

[25] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *Speech and Audio Processing, IEEE Transactions on,* vol. 11, no. 5, pp. 466-475, 2003.

[26] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, “Speech recognition with primarily temporal cues,” *Science,* vol. 270, no. 5234, pp. 303-304, 1995.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, no. 7, pp. 2125-2136, 2011.

[28] J. Ma, Y. Hu, and P. C. Loizou, “Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions,” *The Journal of the Acoustical Society of America,* vol. 125, pp. 3387, 2009.

[29] F. Chen, and P. C. Loizou, “Predicting the intelligibility of vocoded and wideband Mandarin Chinese,” *The Journal of the Acoustical Society of America,* vol. 129, pp. 3281, 2011.

[30] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 27, no. 2, pp. 113-120, 1979.

[31] Y. Lu, and P. C. Loizou, “A geometric approach to spectral subtraction,” *Speech communication,* vol. 50, no. 6, pp. 453-466, 2008.

[32] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, “Adaptive β-order generalized spectral subtraction for speech enhancement,” *Signal Processing,* vol. 88, no. 11, pp. 2764-2776, 2008.

[33] J. Chen, J. Benesty, Y. Huang, and S. Doclo, “New insights into the noise reduction Wiener filter,” *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 14, no. 4, pp. 1218-1234, 2006.

[34] E. Plourde, and B. Champagne, “Auditory-based spectral amplitude estimators for speech enhancement,” *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 16, no. 8, pp. 1614-1623, 2008.

[35] R. J. McAulay, and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter, ” *Acoustics, Speech and Signal Processing,*

IEEE Transactions on, vol. 28, no. 2, pp. 137-145, 1980.

[36] Y. Tsao, and Y. H. Lai, "Generalized maximum a posteriori spectral amplitude estimation for speech enhancement," *Speech Communication,* vol. 76, pp. 112-126, 2016.

[37] Y. Ephraim, and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on,* vol. 32, no. 6, pp. 1109-1121, 1984.

[38] Y. Ephraim, and H. L. Van Trees, "A signal subspace approach for speech enhancement," *Speech and Audio Processing, IEEE Transactions on,* vol. 3, no. 4, pp. 251-266, 1995.

[39] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *The Journal of Machine Learning Research,* vol. 11, pp. 3371-3408, 2010.

[40] J. Martens, "Deep learning via Hessian-free optimization," *in Proc. International Conference on Machine Learning*, pp. 735-742, 2010.

[41] P. C. Loizou, "Introduction to cochlear implants," *Engineering in Medicine and Biology Magazine, IEEE,* vol. 18, no. 1, pp. 32-42, 1999.

[42] M. F. Dorman, P. C. Loizou, and D. Rainey, "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," *The Journal of the Acoustical Society of America,* vol. 102, no. 4, pp. 2403-2411, 1997.

[43] F. Chen, and A. H. Lau, "Effect of vocoder type to Mandarin speech recognition in cochlear implant simulation," *in Proc. ISCSLP*, pp. 551-554, 2014.

[44] Y. H. Lai, Y. Tsao, and F. Chen, "Effects of adaptation rate and noise suppression on the intelligibility of compressed-envelope based speech," *PLoS ONE*, pp. 10.1371/journal.pone.0133519, 2015.

[45] F. G. Zeng, "Trends in cochlear implants," *Trends in amplification,* vol. 8, no. 1, pp. 1-34, 2004.

[46] N. A. Whitmal, S. F. Poissant, R. L. Freyman, and K. S. Helfer, "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," *The Journal of the Acoustical Society of America,* vol. 122, no. 4, pp. 2376-2388, 2007.

[47] B. Roberts, R. J. Summers, and P. J. Bailey, "The intelligibility of noise-vocoded speech: spectral information available from across-channel comparison of amplitude envelopes," *Proceedings of the Royal Society of London B: Biological Sciences,* vol. 278, no. 1711, pp. 1595-1600, 2011.

[48] P. C. Loizou, *Speech enhancement: theory and practice*: CRC press, 2013.

[49] L. L. Wong, S. D. Soli, S. Liu, N. Han, and M.-W. Huang, "Development of the Mandarin hearing in noise test (MHINT)," *Ear and hearing,* vol. 28, no. 2, pp. 70-74, 2007.

[50] K. Nie, G. Stickney, and F. G. Zeng, "Encoding frequency modulation to improve cochlear implant performance in noise," *Biomedical Engineering, IEEE Transactions on,* vol. 52, no. 1, pp. 64-73, 2005.

[51] G. S. Stickney, F. G. Zeng, R. Litovsky, and P. Assmann, "Cochlear implant speech recognition with speech maskers," *The Journal of the Acoustical Society of America,* vol. 116, no. 2, pp. 1081-1091, 2004.

[52] H. J. Steeneken, and T. Houtgast, "A physical method for measuring speech-transmission quality," *The Journal of the Acoustical Society of America,* vol. 67, no. 1, pp. 318-326, 1980.

[53] R. L. Goldsworthy, and J. E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *The Journal of the Acoustical Society of America,* vol. 116, no. 6, pp. 3679-3689, 2004.

[54] A. ANSI, "S3. 5-1997, Methods for the calculation of the speech intelligibility index," *New York: American National Standards Institute*, 1997.

[55] F. Chen, and P. C. Loizou, "Predicting the intelligibility of vocoded speech," *Ear and hearing,* vol. 32, no. 3, pp. 331, 2011.

[56] C. B. Hicks, and A. M. Tharpe, "Listening effort and fatigue in school-age children with and without hearing loss," *Journal of Speech, Language, and Hearing Research,* vol. 45, no. 3, pp. 573-584, 2002.

[57] R. H. Gifford, J. K. Shallop, and A. M. Peterson, "Speech recognition materials and ceiling effects: Considerations for cochlear implant programs," *Audiology and Neurotology,* vol. 13, no. 3, pp. 193-205, 2008.

[58] G. A. Studebaker, "A rationalized arcsine transform," *Journal of Speech, Language, and Hearing Research,* vol. 28, no. 3, pp. 455-462, 1985.

[59] Y. Hu, and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 16, no. 1, pp. 229-238, 2008.

[60] S. Rangachari, and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech communication,* vol. 48, no. 2, pp. 220-231, 2006.

[61] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters, IEEE,* vol. 6, no. 1, pp. 1-3, 1999.

[62] I. Cohen, and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *Signal Processing Letters, IEEE,* vol. 9, no. 1, pp. 12-15, 2002.

[63] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on,* vol. 9, no. 5, pp. 504-512, 2001.

[64] A. Bhattacharya, and F. G. Zeng, "Companding to improve cochlear-implant speech recognition in speech-shaped noisea)," *The Journal of the Acoustical Society of America,* vol. 122, no. 2, pp. 1079-1089, 2007.

[65] L. Turicchia, and R. Sarpeshkar, “A bio-inspired companding strategy for spectral enhancement,” *IEEE transactions on speech and audio processing,* vol. 13, no. 2, pp. 243-253, 2005.

[66] A. J. Oxenham, A. M. Simonson, L. Turicchia, and R. Sarpeshkar, “Evaluation of companding-based spectral enhancement using simulated cochlear-implant processing,” *The Journal of the Acoustical Society of America,* vol. 121, no. 3, pp. 1709-1716, 2007.

[67] L. M. Friesen, R. V. Shannon, D. Baskent, and X. Wang, “Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants,” *The Journal of the Acoustical Society of America,* vol. 110, no. 2, pp. 1150-1163, 2001.

[68] Q. J. Fu, R. V. Shannon, and X. Wang, “Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing,” *The Journal of the Acoustical Society of America,* vol. 104, no. 6, pp. 3586-3596, 1998.

[69] M. F. Dorman, P. C. Loizou, and D. Rainey, “Simulating the effect of cochlear-implant electrode insertion depth on speech understanding,” *The Journal of the Acoustical Society of America,* vol. 102, no. 5, pp. 2993-2996, 1997.

[70] M. J. Goupell, P. Majdak, and B. Laback, “Median-plane sound localization as a function of the number of spectral channels using a channel vocoder,” *The Journal of the Acoustical Society of America,* vol. 127, no. 2, pp. 990-1001, 2010.

[71] F. Chen, “The relative importance of temporal envelope information for intelligibility prediction: A study on cochlear-implant vocoded speech,” *Medical engineering & physics,* vol. 33, no. 8, pp. 1033-1038, 2011.